DOCUMENT RESUME

ED 378 196 TM 022 526

AUTHOR Tang, Huixing
TITLE Step Fit Analysis with Polytomously Scored Items.
PUB DATE Apr 94
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *Goodness of Fit; *Item Response Theory; Models; *Scores; Scoring; Simulation; *Test Items
IDENTIFIERS Performance Based Evaluation; *Polytomous Scoring; *Step Fit Analysis

ABSTRACT

Fit analysis is widely performed in item response theory (IRT) based test development to assess the fit of individual items to the IRT model being used. The paper explores a step fit analysis procedure that is an extension of IRT-based item fit diagnostics applied to the response categories present in popular performance-based tasks. The step fit procedure involves computing category fit statistics and constructing category fit plots. Category fit statistics are used to flag possible misfits at the response category level, and category fit plots are used to facilitate investigation of potential fit problems by displaying the magnitude and pattern of deviations for each step of a polytomous item. A step-by-step description of the procedure is presented, with findings from a simulation study and from a real data application involving 8 to 11 tasks for 5 content areas at 11 grade levels, with 400 to 1,000 examinees (students) at each grade level. Results suggest that the procedure is promising for examining the statistical characteristics at the response category level and for diagnosing potential problems in item content at the category level, while clarifying and improving the scoring rubric. Three tables and six figures present study results. (Contains 4 references.) (SLD)

Step Fit Analysis with Polytomously Scored Items

by

Huixing Tang

The Psychological Corporation

Paper presented at the 1994 annual meeting of the
American Educational Research Association
New Orleans, Louisiana

I. RATIONALE

Fit analysis is widely used in IRT-based test development. It is usually performed at the item level to evaluate the fit of individual items to the IRT model being used. It is often used to assess item consistency (the extent to which an item is consistent with other items in measuring examinee ability), or interitem dependency (the extent to which the items are statistically dependent for examinees of the same ability). Such information is important for test developers in the selection, revision or analysis of individual items.

While fit analysis at the item level is useful, it may be insufficient for examining the statistical characteristics of performance-based items or tasks which typically involve multiple steps and are scored on a polytomous scale. For such tasks, item fit statistics may serve as an index of 'global' fit, but they do not provide information as to where, or at which step(s), the misfit occurs. In addition, such global indicators do not report the extent of misfit in terms of the number of steps involved. Fit analysis at each step of the performance task is needed, particularly when the primary purpose of the fit analysis is to revise the item or to adjust the scoring rubric.

The purpose of the current investigation is to explore a step or category fit analysis procedure, which is an extension of IRT-based item fit diagnostics applied to the response categories present in popular performance-based tasks. The step fit procedure involves (a) computing category fit statistics, and (b) constructing category fit plots. The category fit statistics are used to flag possible misfit at the response category level, in much the same way as item fit statistics are used to flag possible misfit at the item level. The category fit plots, on the other hand, are used to facilitate investigation of potential fit problems by displaying the magnitude and the pattern of deviations for each step of a polytomous item. A step-by-step description of the procedure is presented, followed by discussion of findings from a simulation study and a real data application. The paper concludes with a summary of the strengths and the limitations of the procedure.

II. THE PROCEDURE

Mean Square Category Fit Statistics

Wright and Masters (1982) describe two mean square fit statistics: infit mean square and outfit mean square. Infit mean square is information weighted and outfit mean square is sensitive to outliers. When data fit the model, both statistics have an expectation of 1. A value considerably larger than 1 may suggest inconsistency, while a value considerably smaller than 1 may indicate item dependency.

BIGSTEPS, a popular Rasch model computer program (Linacre & Wright 1993), also provides mean square fit statistics at the category score level. These statistics are computed by averaging, with or without weighting by item score variance, the residuals only across observations within each score category. For example, the mean squares for score category 1 are computed by averaging the residuals only across the examinees who scored in category 1. At both the item and the category levels, the same item score residuals are used for computation. The mean

squares computed in this way are really *item* fit indices at the category level. They describe how well the item fits the model for examinees at each score category.

The current study explores an alternative way to compute mean square fit statistics at the category level. First, the *category* standardized residuals are generated. These residuals are the standardized differences between the observed and the expected category scores. The category standardized residuals are then averaged over examinees in all score categories to compute mean square category fit statistics. These mean squares can be interpreted as category fit indices which describe how well the *category* fits the model across all examinees. Those who score in a particular category are expected to have a high probability of scoring in that category. Those who do not score in that category are expected to have a low probability of scoring in the category. The category misfits the model to the extent the observed responses depart from model expectation.

The computation of the new category fit statistics is a simple extension of the computation of item mean squares to the category level. They are computed as follows:

1. Calibrate the data with the Rasch partial credit model (Masters, 1982; Wright & Masters, 1982) to obtain B_n, the ability estimate for person n, and D_{ij}, the difficulty of step j of item i. There are m step difficulty estimates for an item with $m+1$ score categories.

2. Using the estimates obtained in step 1, compute P_{nik}, the probability of person n scoring in the kth category on item i by

$$P_{nik}=\frac{\exp\sum_{j=0}^{k}(B_n-D_{ij})}{\sum_{k=0}^{m_i}\exp\sum_{j=0}^{k}(B_n-D_{ij})}$$

where $k= 0,1,..,m_i$, and $\sum_{j=0}^{0}(B_n-D_{ij})=0$.

3. Create a person by category response matrix using the item scores. For each score category, an examinee is assigned a score of 1 if the score falls in that category, or 0 if the score falls in any other category (see Table 1) .

Table 1. A person by category score matrix

n (Person)	X_{ni} (Item score)	X_{nik} (Category Score) 0	1	2	K
1	0	1	0	0	0
2	1	0	1	0	0
3	2	0	0	1	0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
N	K	0	0	0	.

4. Using the dichotomous category score, X_{nik}, as the category 'observed' score and P_{nik} as the category expected score, compute Z_{nik}, the standardized category score residual by

$$Z_{nik}=\frac{X_{nik}-P_{nik}}{\sqrt{W_{nik}}}$$

where $W_{nik}=P_{nik}(1-P_{nik})$, the variance of X_{nik}.

5. Compute U_{ik}, the unweighted category mean square for category *k* on item *i* by

$$U_{ik}=\sum_{n=1}^{N} Z_{nik}^2/N$$

and V_{ik}, the weighted category mean square by

$$V_{ik}=\sum_{n=1}^{N} W_{nik}*Z_{nik}^2/\sum_{n=1}^{N} W_{nik}$$

As shown in the foregoing, the category mean squares are computed in the same way as item mean squares for dichotomously scored items (Wright & Stone, 1979), and are intended for the same type of interpretation. Just as item fit statistics describe the extent to which an item is consistent with the set of items as a whole in rank-ordering examinees, category fit statistics indicate the extent to which a particular response category is functioning consistently with the test to measure examinee ability.

Category Fit Plot

The purpose of generating the category fit plot is to help interpret the category fit statistics by displaying the pattern as well as the magnitude of the deviations. Two types of plots are proposed: (1) the expected and the observed proportion plot, and (2) the residual plot.

The observed and the expected proportion plot is constructed as follows:

1. Divide the examinees into K ability groups using equal-distance interval method, equal-frequency method, or total raw score categories.

2. Compute E_k, the expected proportion of examinees in the *k*th group scoring in each category by averaging the probabilities of the examinees in the group responding in that category.

3. Compute O_k, the observed proportion of examinees in the *k*th group scoring in each category.

4. Compute SE_k, the standard error of the expected proportion for the *k*th group by

$$SE_k = \sqrt{\frac{E_k(1-E_k)}{N_k}}$$

5. Compute the upper and the lower error bound for O_k (e.g., +2 and -2 times SE_k).

6. Plot E_k vs O_k and the upper and lower error bounds for each category across the intervals.

The residual plot is constructed by following steps 1 through 3 listed above. Then, for each ability group or interval, the residual proportion is computed by subtracting the expected proportion from the observed proportion. The residual proportions are then plotted for each category across the intervals.

III. A SIMULATION STUDY

Design of the Simulation Study

A simulation study was conducted to explore the statistical properties of the category mean squares in relation to examinee sample size and the number of score categories in an item. The examinee sample size varies from 500 to 900 by increments of 200. The number of item score categories varies from 4 to 8 by increments of 2. Each calibration set consists of nine 'null' (non-misfitting) items and one misfitting item. 100 replications were conducted for each sample size and each number-of-score-category level.

Both the examinee ability and step difficulty estimates were simulated using the standard normal distribution. The step difficulties were randomly generated for each replication to control for the effect of the step difficulty on model-data fit. The misfitting item was simulated using a

secondary ability distribution imperfectly correlated (r=.6) with the primary ability distribution used to simulate the remainder of items.

Findings

Tables 2 and 3 present, for each score category, the mean of the mean square statistics over replications (and over items for the null items), and the percentage of replications in which the category is flagged as misfitting (mean square value greater than or equal to 1.2 or less than .8). The following observations, among others, are notable:

1. For the null items, the means for the infit (weighted) mean squares are close to 1 and the percentage of misfitting replications is below 1 across all categories. This finding suggests that the infit category mean square under the null situation is sensitive to neither the sample size nor the number of response categories in an item.

2. For the misfitting item, neither the infit mean square nor the percentage of misfitting replications is sensitive to sample size, but both appear to be sensitive to the number of score categories. As the number of score categories increases, both the mean square and the percentage of misfits increase (if we compare the values for the first and the last score category across different number-of-category levels). The magnitude of the sensitivity is, however, small.

3. For outfit (unweighted) category mean squares, the mean square values and the percentage of misfits are sensitive to the number of score categories for both the null and the misfitting items. The mean squares become increasingly smaller than 1 for the null items and increasingly greater than 1 for the misfitting item as the number of categories increases. The percentage of misfits is high even for the null items and rapidly increases as the number of categories increases. They are not, however, sensitive to sample size in both the null and the misfitting situations.

4. Except for the infit mean squares in the null situation, the percentage of misfits is far greater in the extreme score categories than in the middle categories.

One major limitation of the category mean squares is the excessive sensitivity of the unweighted category mean squares to the number of categories in an item. Recall that the unweighted category mean square is an unweighted average of Z_{nik}^2's. It is computed by

$$U_{ik}=\frac{1}{N}\sum_{n=1}^{N} Z_{nik}^2=\frac{1}{N}\sum_{n=1}^{N} \frac{(X_{nik}-P_{nik})^2}{P_{nik}(1-P_{nik})}$$

and

$$\sum_{k=0}^{K} P_{nik}=1$$

As K increases, P_{nik} decreases, so does the variance of X_{nik}. It can be shown that, as K increases and the category score variance decreases, Z_{nik} will become increasingly small for X_{nik}=0, and increasingly large for X_{nik}=1. The more extreme Z_{nik} becomes, the more likely it is for U_{ik} to be either considerably larger or smaller than 1.

Table 2. Mean of Infit Category Mean Squares and % Misfitting

		Score Category													
		0		1		2		3		4		5		6		7	
Category	N	Mean	%	Mean	%	Mean	%	Mean	%	Mean	%	Mean	%	Mean	%	Mean	%
Null																	
3	500	.98	.1	.99	0	.99	0	.98	.1								
	700	.98	.0	.99	0	.99	0	.98	0								
	900	.98	.0	.99	0	.99	0	.98	0								
5	500	.98	.3	.98	0	.99	0	.99	0	.99	0	.97	.4				
	700	.97	.3	.98	0	.99	0	.99	0	.99	0	.98	.2				
	900	.97	.2	.99	0	.99	0	.99	0	.99	0	.98	0				
7	500	.97	.8	.98	0	.99	0	.99	0	.99	0	.99	0	.98	0	.98	1
	700	.97	.2	.98	0	.99	0	.99	0	.99	0	.99	0	.98	0	.98	.3
	900	.97	.2	.98	0	.99	0	.99	0	.99	0	.99	0	.98	0	.97	.1
Misfitting																	
3	500	1.2	75	1.1	0	1.1	2	1.2	68								
	700	1.2	78	1.1	0	1.1	2	1.2	69								
	900	1.2	75	1.1	0	1.1	1	1.2	72								
5	500	1.3	84	1.1	2	1.1	0	1.1	0	1.1	3	1.3	77				
	700	1.3	82	1.1	3	1.1	0	1.1	0	1.1	2	1.3	76				
	900	1.3	81	1.1	3	1.1	0	1.1	0	1.1	2	1.3	75				
7	500	1.3	86	1.1	9	1.1	0	1.1	0	1.1	0	1.1	0	1.1	6	1.3	80
	700	1.3	87	1.1	7	1.1	1	1.1	0	1.1	0	1.1	0	1.1	4	1.3	77
	900	1.3	86	1.1	6	1.1	1	1.1	0	1.1	0	1.1	0	1.1	3	1.3	82

Table 3. Mean of Outfit Category Mean Squares and % Misfitting

		Score Category							
		0	1	2	3	4	5	6	7
Category	N	Mean %	Mean %	Mean %	Mean %	Mean %	Mean %	Mean %	Mean %
Null									
3	500	.87 40	.95 5	.96 3	.87 41				
	700	.88 37	.95 2	.96 2	.86 37				
	900	.88 36	.95 2	.96 2	.87 35				
5	500	.77 74	.85 45	.91 14	.92 14	.87 40	.73 75		
	700	.77 73	.85 41	.91 11	.92 10	.86 37	.73 73		
	900	.76 74	.85 37	.91 8	.92 9	.86 34	.75 70		
7	500	.68 84	.74 73	.83 50	.87 35	.87 34	.83 49	.74 73	.65 88
	700	.67 85	.75 73	.82 50	.87 31	.87 30	.83 47	.74 71	.65 87
	900	.67 85	.75 72	.83 46	.87 25	.87 27	.83 43	.75 72	.66 86
Misfitting									
3	500	2.2 99	1.4 51	1.3 51	2.1 99				
	700	2.2 99	1.3 50	1.3 47	2.2 99				
	900	2.2 99	1.4 54	1.3 49	2.2 99				
5	500	3.4 97	2.8 95	1.7 93	1.7 80	2.5 91	3.5 99		
	700	3.6 98	2.7 95	1.7 93	1.7 91	2.5 95	3.6 99		
	900	3.7 97	2.7 94	1.7 97	1.7 95	2.5 95	3.6 99		
7	500	4.1 99	3.5 96	3.1 97	2.4 96	2.1 96	2.9 96	3.7 99	4.2 99
	700	4.5 99	3.7 99	3.0 98	2.5 97	2.3 98	3.0 99	3.8 99	4.3 99
	900	4.5 99	3.9 99	3.1 99	2.5 99	2.3 99	3.0 99	3.9 99	4.5 99

VI. A REAL-DATA APPLICATION

The step fit procedure has been applied to a large pool of try-out performance-based tasks scored on a 4 or 5 point scale (0-3 or 0-4) developed at The Psychological Corporation. These tasks cover five content areas in eleven grade levels and have been administered to a national sample of elementary and high school students. Each form consisted of 8 to 11 tasks for each content area and was administered to 400 to 1000 examinees. The purpose of this application is to examine the sensitivity of the category fit statistics with real data, and to illustrate how the category fit plots can help in the interpretation of the fit statistics.

Sensitivity of the category fit statistics

As item and category mean squares are indices for the same performance characteristics at different score levels, they are expected to function consistently in identifying model-data misfit. To examine the consistency between the item and the category mean squares and their relative sensitivity, the percentage of items with at least one category flagged for misfit was compared with the percentage of items considered misfit at the item level. An item or a category is flagged for misfit if the mean square value is less than .8 or greater than or equal to 1.2. Figure 1 presents the number and the percentage of items in each of the four classification categories: misfit for (1) neither the item nor the category, (2) the category only, (3) the item only, and (4) both the item and the category.

For infit mean squares, the item and the category fit values are quite consistent with each other. Less than five percent of the items fall in category 2 or 3, where the item and category do not agree in rejection. For outfit mean squares, however, there appears to be a great deal of discrepancy between the item and the category fit statistics in flagging the misfitting status. 61.5 percent of the items not flagged for misfit at the item level had at least one misfitting category. On the other hand, only 2 items (.2%), flagged for misfit at the item level, were not flagged for misfit at any of the score categories. This suggests that the unweighted mean squares are too sensitive when the same rejection criteria as those for the weighted mean square values are used.

Figure 2 presents the relative frequency distributions of item and category mean square statistics. For this plot, only those items (n=921) scored on the four point (0 to 3) scale were used so that 0 and 3 were the extreme score categories for all the items. For the weighted mean squares, the relative frequency distributions for categories 0 and 3 are quite close to the distribution for the item, except that they are a bit more leptokurtic (peaked) than the distribution for the item mean squares, particularly for category 3. On the other hand, the relative distributions for the two middle categories are considerably more peaked; hardly any items fall in the rejection regions. This could be due to the fact the extreme categories are far more discriminative than the middle categories and are therefore more likely to produce high residual values.

Item Fit

Step Fit		Yes	No	
Yes		923	52	n
		86.7	4.9	%
No		46	43	n
		4.3	4.1	%

Infit

Item Fit

Step Fit		Yes	No	
Yes		203	2	n
		19.1	.2	%
No		655	204	n
		61.5	19.2	%

Outfit

Figure 1. Relative sensitivity of item and category fit statistics

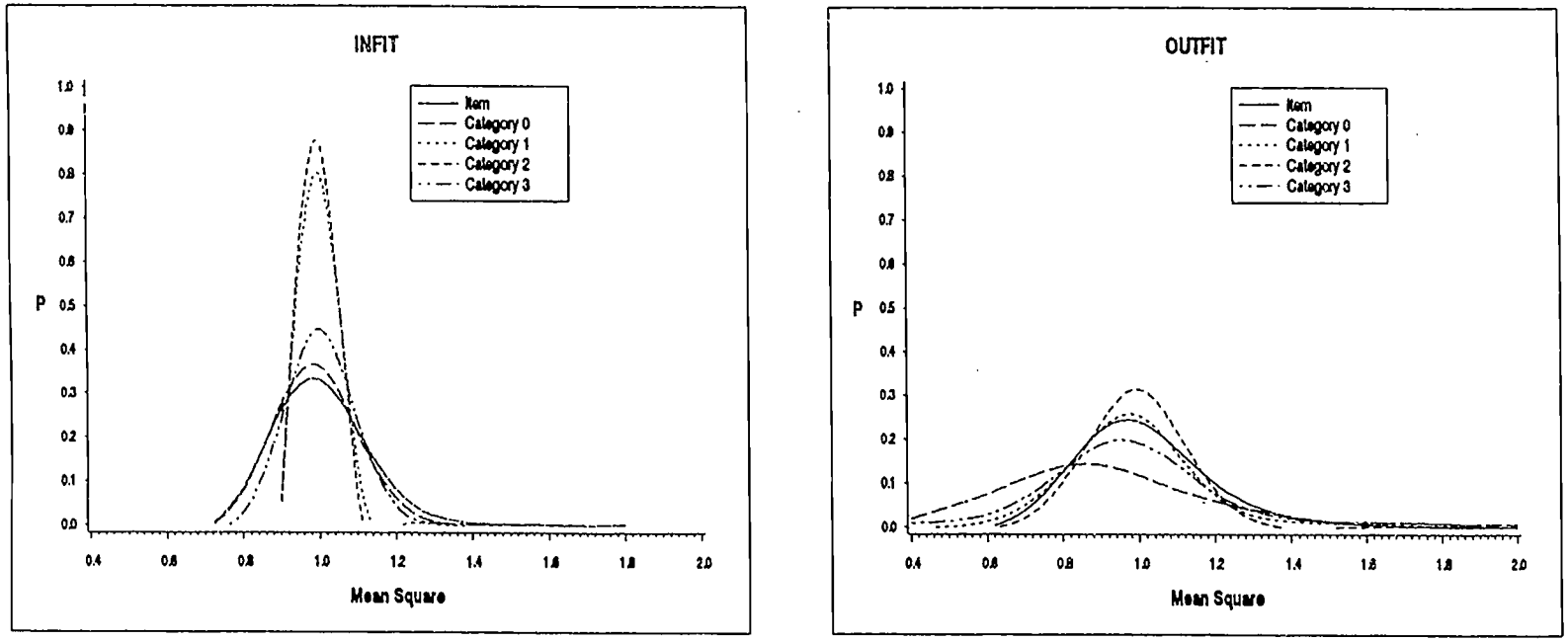


Figure 2. The relative frequency distribution of item and category mean squares.

The unweighted mean square distributions are, in general, considerably more platykurtic (flat) than the weighted mean square distributions. The distributions for the two extreme categories are more flat than the distribution for the item. This indicates that the unweighted category mean squares are more sensitive than both the weighted category mean squares and the item mean squares (weighted or unweighted).

Using category fit plots to help interpret the fit statistics

The category fit plots may enhance the interpretation of fit statistics. This feature will be illustrated using the category fit plots for three second grade Language items, all extracted from the same calibration set. One item was scored on a four point (0 to 3) scale, and the other two on a five point (0 to 4) scale. Figures 3 through 5 show the expected and observed proportion category fit plots for these items. These plots were constructed by first rank-ordering examinees in 20 groups according to their ability estimates. The expected and the observed proportions in each response category were then computed for each of the 20 groups. The error bounds for these plots were constructed using plus and minus 1 standard error of the observed proportion for each group.

Figure 3 shows an item that fits quite well in all categories. The infit values are all close to 1. All the observed proportion curves fall within the error bound. The outfit values are too high for category 3 (1.57) and too low for category 4 (.77). But the observed proportion curve in neither category exhibits any significant departure from the expected proportion curve. As noted earlier, the outfit values could be spuriously high or low due to extremely small variances of the category score.

Figure 4 shows a very unusual item. The item infit and outfit values are 1.36 and 1.56, respectively, which suggest that this item is inconsistent with the test as a whole in measuring examinee ability. The examinee performance shown in each category clearly illustrates the inconsistency. For category 0, both the expected and the observed proportions are well below .1 across all groups. Although both infit and outfit values fall within the acceptable range, this category is not really discriminating among examinees. Category 2, on the other hand, is the most probable response category for almost all the ability groups. The misfit pattern is quite obvious. The observed proportion curve is considerably more 'flat' than the expected proportion curve. And the observed proportion is constantly high across all ability groups. Even for the highest ability group, the observed proportion for scoring in this category is nearly .6, almost .4 (or 40 percent) higher than is expected by their overall test performance. Categories 2 and 3, the two higher score categories, exhibit similar misfit patterns: Lower proportions of the high ability groups scored in these categories than expected, while higher proportions of the low ability groups scored in these categories.

Figure 5 shows a different kind of misfit. The item infit and outfit values are .79 and .71, respectively, suggesting possible item dependency. One commonly observed fact associated with inter-item dependency is that the high ability examinees tend to score higher than expected and/or low ability examinees tend to score lower than expected. Examining the plots for categories 0 and 4, the two most discriminative categories, we observe that for category 0, consistently higher proportions of the low ability groups scored 0 while consistently lower proportions of the high

ability groups scored 0. For category 4, on the other hand, lower proportions of the low ability groups (from group 4 to 10) scored 4, while higher proportions of the three highest ability groups (groups 18 through 20) scored in this category. The deviation patterns exhibited in these two plots indicate that the high ability examinees, in general, are doing better while the low ability examinees are doing worse on this item than expected by their overall test performance.

Figure 6 presents the residual plots for all three items discussed in this section. They provide an alternative way to examine the magnitude and the pattern of residuals. Figure 6A shows that all the residual curves cluster tightly around 0. Figures 6B and 6C reveal different patterns of the residual plots for different score categories as well as the magnitude of the residuals across ability intervals. The residual plot for category 1 of item 4, for example, is consistently below 0 for most of the lower ability groups and is consistently above 0 for the five highest ability groups. Among these high ability groups, the magnitude of the residuals increases as the ability increases.

V. SUMMARY AND CONCLUSION

This study explored a step fit procedure using category fit statistics and category fit plots. The category fit statistics are a simple extension of the item mean square fit statistics to the category score level. The weighted category fit statistics are as sensitive as and quite consistent with the item fit statistics in identifying model-data misfit. The unweighted category fit statistics, on the other hand, are too sensitive to be practically useful for category fit analysis. The excessive sensitivity of the unweighted category mean square is largely a function of its sensitivity to the number of categories in the item The category fit plots may enhance the interpretation of fit statistics by displaying the pattern and the magnitude of the deviations at the category level.

As educational testing tends toward performance-based tasks with polytomous scoring, statistical analyses at the item-step level become increasingly important. The procedure explored in this study looks beyond the item level into the more fundamental units that comprise the item. This could be a promising procedure not only for examining the statistical characteristics at the response category level, but also, when integrated with substantive analysis, for diagnosing potential problems in item content at the category level and for clarifying and improving the scoring rubric.

References:

Linacre, M. J. & Wright, B. D. (1993). A user's guide to BIGSTEPS. Chicago: MESA Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrica, 47(2), 149-174.

Wright, B. D. & Masters, G. N. (1982). Rating scale analysis: Rasch measurement. Chicago: MESA Press.

Wright, B. D. & Stone, M. H. (1979). Best test design. Chicago: MESA Press.

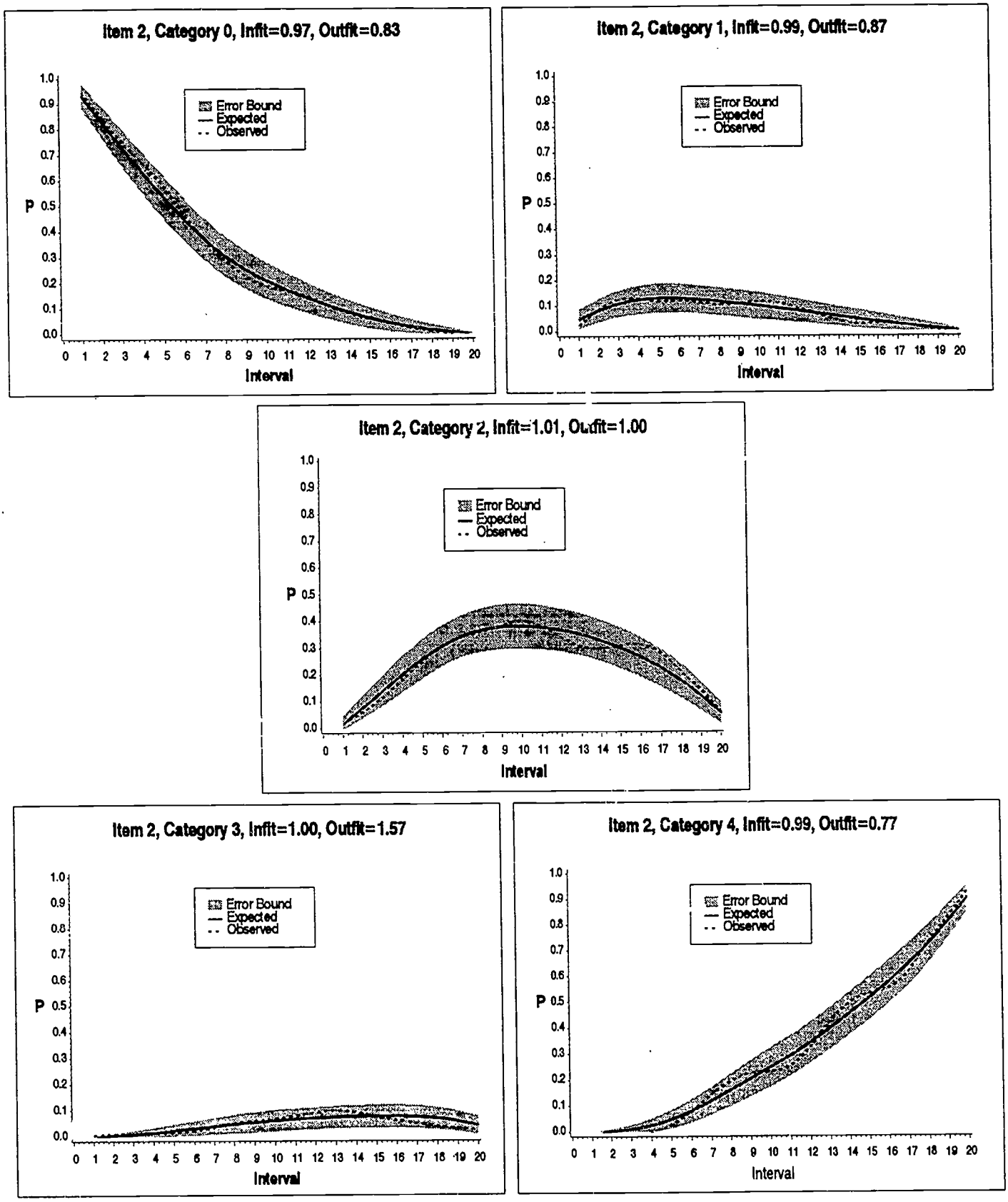


Figure 3. Category fit plots for item 2

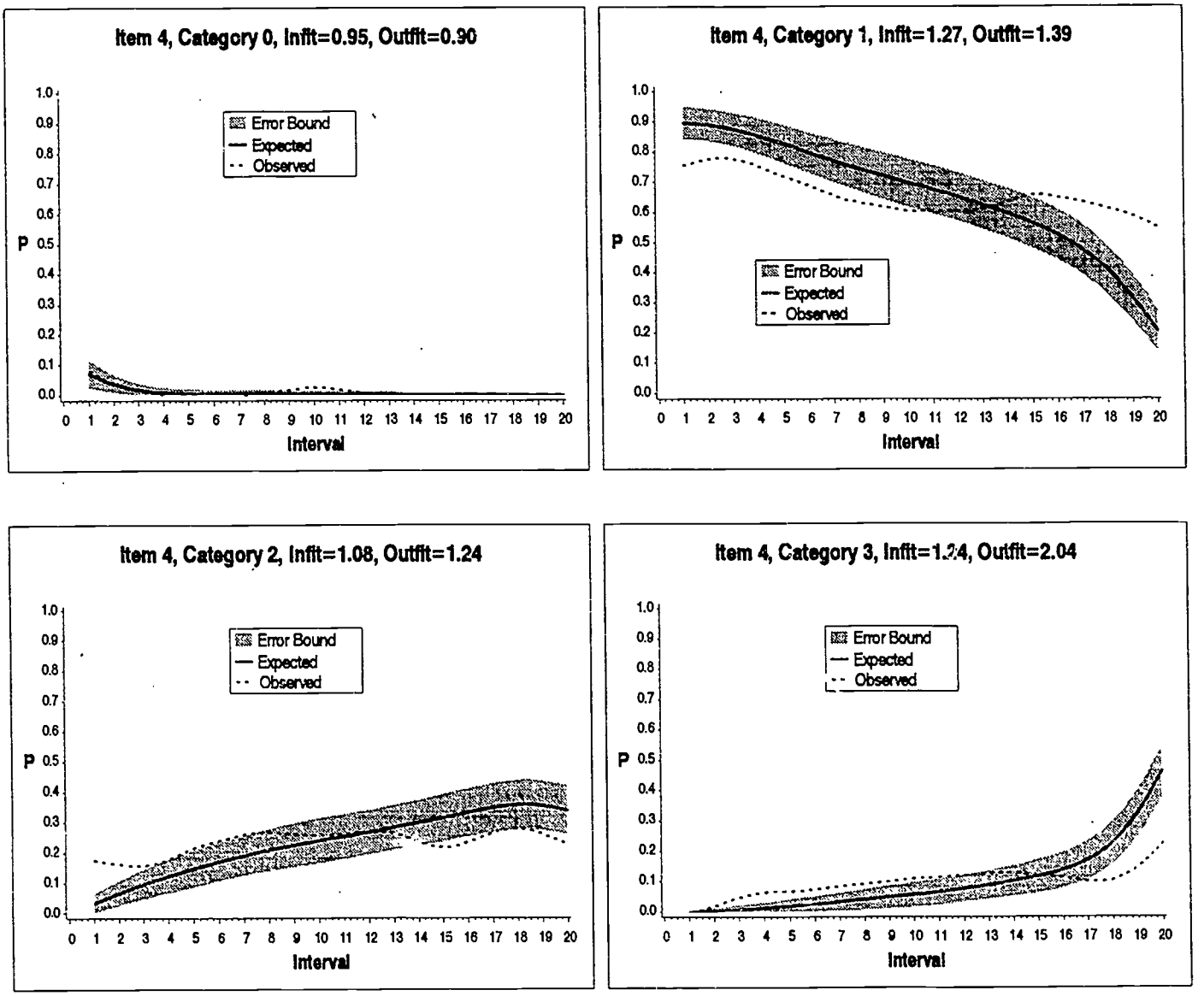


Figure 4. Category fit plots for item 4

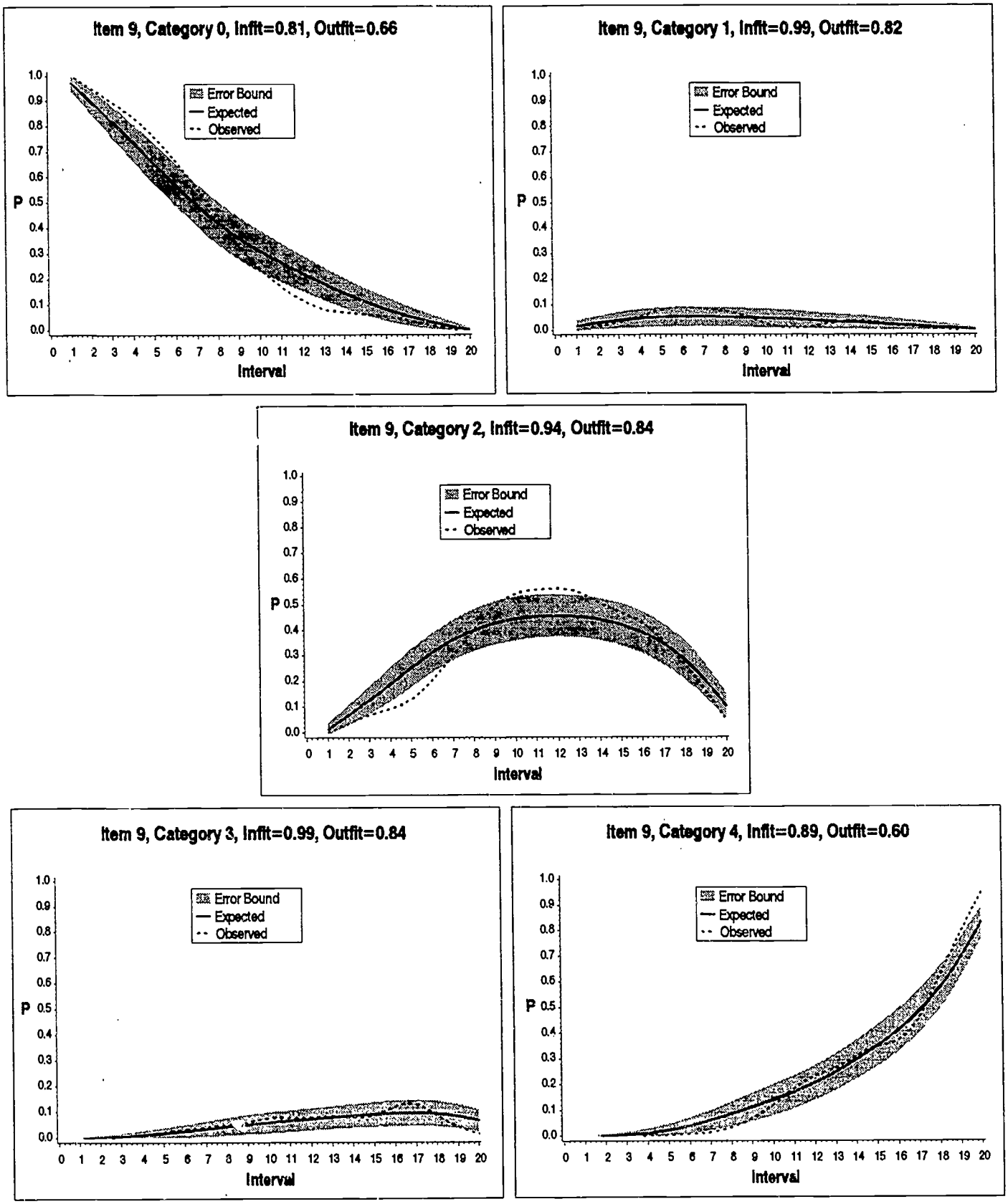
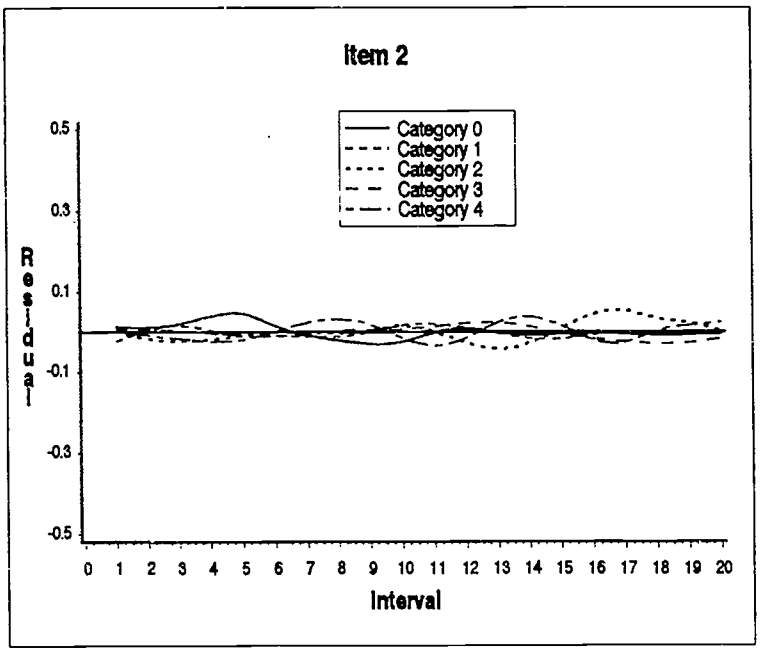
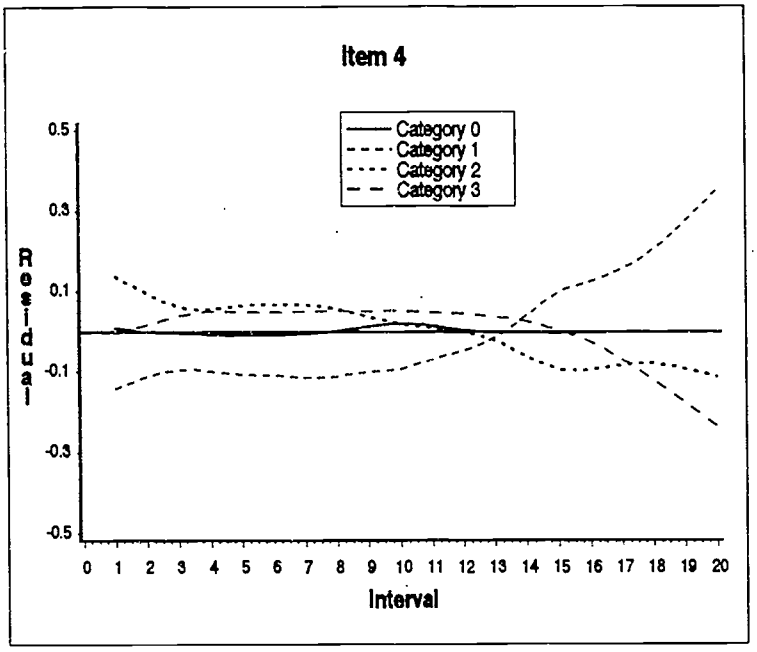


Figure 5. Category fit plots for item 9

6A



6B



6C

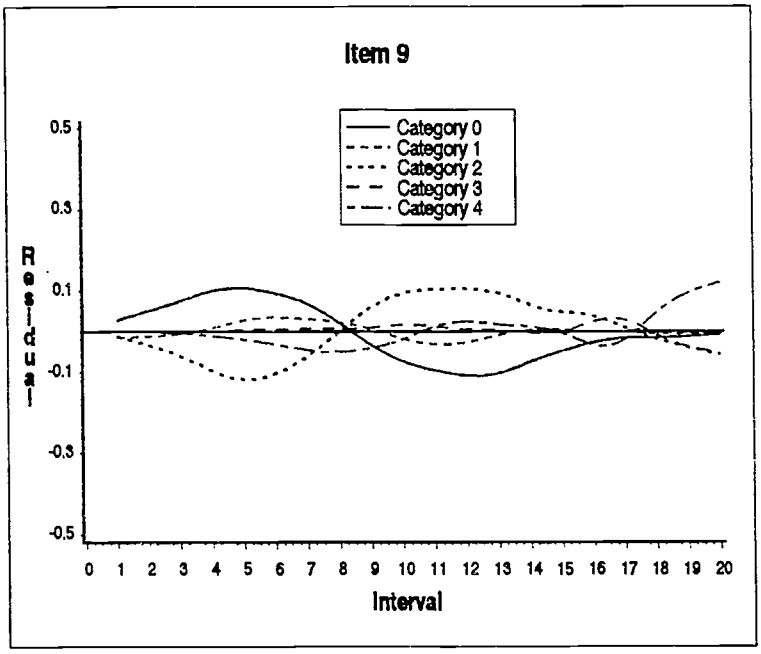


Figure 6. The residual category fit plot for items 2, 4 and 9